Обучение машинного перевода без параллельных текстов*

Ярошенко А. М. Бахтеев*[1] *О. Ю. Стрижов*[2] *В. В.
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В данной работе исследуется задача машинного перевода между двумя языками. Для решения часто используются параллельные предложения, то есть совпадающие по смыслу фразы на двух языках. В работе рассматривается альтернативная модель, не требующая большого количества параллельных предложений. Она состоит из двух нейросетей типа Seq2Seq, имеющих общее скрытое пространство. [Тут добавится что-то от меня]. Для проверки качества модели проводится вычислительный эксперимент по переводу предложений между близкими языками, такими как русский и украинский.

Ключевые слова: *нейронные сети, машинный перевод, автокодировщики.*

Введение

В зависимости от специфики пары языков выделяют несколько подходов к машинному переводу. При наличии достаточного числа параллельных предложений(порядка миллиона) использование глубоких нейронных сетей привело к тотальному коллапсу([1],[2]).

Но для многих пар языков нет достаточной базы примеров. Одним из подходов на основе параллельных предложений является пополнение обучающей выборки переводами с предыдущих итераций работы нейронной сети([3]). Ниже представлено решение задачи машинного перевода при отсутствии достаточного количества параллельных предложений ([4], [5], [6]).

Такой подход был использован в [7] для пары языков французский-русский. В данной работе будет проведен схожий эксперимент для перевода с русского на украинский. В модели используются 2 типа автокодировщиков: LSTM ([8], [9]), которые реализуют перевод слов в скрытое пространство, и сеть-дискриминатор, определяющая по векторному представлению язык исходного предложения. Сети LSTM тренируются так, чтобы представление одного и того же предложения на разных языках совпадало в скрытом пространстве, то есть, чтобы дискриминатору было сложнее определить язык, к которому относится вектор. Обучение состоит из двух фаз. На первой обучается дискриминатор: предложение кодируется с добавлением шума ([11]) и подаётся на вход и происходит перераспределение весов. На второй стадии происходит перераспределение весов уже у сетей-кодировщиков. После обновления весов вычисляется значение функции потерь. Качество переводчика в работе оценивается с помощью метрики BLEU([12]).

Литература

[1] Bilingual word embeddings for phrase-based machine translation / Will Y Zou, Richard Socher, Daniel Cer [и др.] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. С. 1393–1398.

[2] On the properties of neural machine translation: Encoder-decoder approaches / Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau [и др.] // arXiv preprint arXiv:1409.1259. 2014.

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

[3] Bertoldi Nicola, Federico Marcello. Domain adaptation for statistical machine translation with monolingual resources // Proceedings of the fourth workshop on statistical machine translation / Association for Computational Linguistics. 2009. С. 182–189.

[4] Google's neural machine translation system: Bridging the gap between human and machine translation / Yonghui Wu, Mike Schuster, Zhifeng Chen [и др.] // arXiv preprint arXiv:1609.08144. 2016.

[5] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. 2014. С. 3104–3112.

[6] Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua. Neural machine translation by jointly learning to align and translate // arXiv preprint arXiv:1409.0473. 2014.

[7] Lample Guillaume, Denoyer Ludovic, Ranzato Marc'Aurelio. Unsupervised machine translation using monolingual corpora only // arXiv preprint arXiv:1711.00043. 2017.

[8] Gers Felix A, Schmidhuber Jürgen, Cummins Fred. Learning to forget: Continual prediction with LSTM. 1999.

[9] Graves Alex, Schmidhuber Jurgen. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. 2005. Т. 18, № 5-6. С. 602–610.

[10] Goldberg Yoav, Levy Omer. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method // arXiv preprint arXiv:1402.3722. 2014.

[11] Kim Yunsu, Ney Jiahui Geng Hermann. Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder. 2018.

[12] BLEU: a method for automatic evaluation of machine translation / Kishore Papineni, Salim Roukos, Todd Ward [и др.] // Proceedings of the 40th annual meeting on association for computational linguistics / Association for Computational Linguistics. 2002. С. 311–318.